(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
20 December 2001 (20.12.2001)

PCT

(10) International Publication Number
WO 01/96861 A1

(51) International Patent Classification[7]: G01N 33/00, H01J 49/26, 49/40

(21) International Application Number: PCT/SE01/01322

(22) International Filing Date: 12 June 2001 (12.06.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
0002214-5 14 June 2000 (14.06.2000) SE

(71) Applicant and
(72) Inventor: ERIKSSON, Jan [SE/SE]; Studentvägen 15, S-752 34 Uppsala (SE).

(74) Agents: PERNEBORG, Henry et al.; Uppsala Patentbyrå AB, Box 9013, S-750 09 Uppsala (SE).

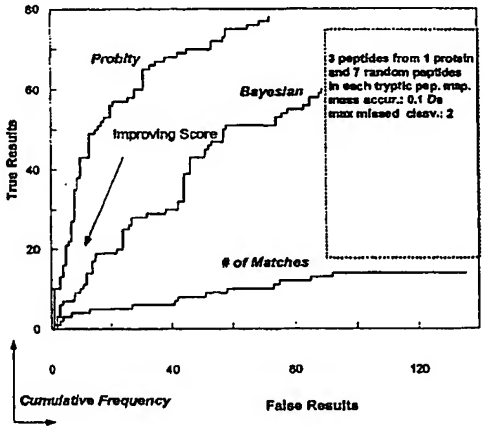
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

**Published:**
— *with international search report*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

*For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

(54) Title: SYSTEM FOR MOLECULE IDENTIFICATION




(57) Abstract: Mass data are typically not unique -I.e., each experimentally determined mass can match randomly one or several molecules in a database. Random matching between mass data and molecules in a database can cause false identification results. In order to minimize false results, random matching must be appropriately accounted for in a method for molecule identification. The invention provides a method to determine, for any molecule in a database and for any experimental and database search constraints, the probability that a particular number of matches between the mass data and masses of molecule constituents results from random matching. The method utilizes the determined probability for random matching to assign scores and rank molecules in a database. The invention further provides a method of generating a frequency function of scores for any experimental condition or database search constraints, wherein the scores relate to random identifications of molecules. Frequency functions are necessary and sufficient tools for testing the significance of a score associated with an identification of an unknown biological molecule.

WO 01/96861 A1

# SYSTEM FOR MOLECULE IDENTIFICATION

## Field of the Invention

The present invention relates to a method and tools for the identification of unknown molecules, and, particularly, a method and tools for molecule identification that provide a solution to the problem of random mass matching.

## Background of the Invention

Identification of a molecule or several molecules in a sample is a technical problem in various fields of research and technology. Molecule identification problems can concern e.g. the tracing of unwanted substances in the environment and the studies of metabolic pathways and disease-state markers in drug development projects. Molecule identification problems can sometimes be solved by the appropriate application of instruments and methods for the acquisition and processing of data from a sample containing the molecules to be identified. One example of data from a sample is mass data. Molecular or molecular constituent mass data can be obtained by a variety of techniques including techniques such as ultra-centrifugation, electrophoresis, and mass spectrometry. Experimental mass data from the sample analyzed is often compared with database-information about known or hypothetical molecules.

In particular, mass spectrometry (MS) combined with database searching has proven to be a useful approach for molecule identification. For example, MS of protein-digests combined with searching in protein and DNA sequence databases is a method of choice for the identification of proteins in *proteomics* projects. The field of proteomics, which include the elucidation of protein function under various cell conditions, is believed to form a future basis for drug design. MS-protein identification involves cleavage of proteins with an enzyme having high digestion specificity (usually trypsin), whereupon the resulting proteolytic products are subjected to mass analysis by either matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) or electrospray ionization mass spectrometry (ESI-MS). The experimentally determined masses are then compared with masses of peptides that individual proteins in a database would yield if they were cleaved by the same enzyme as was used in the experiment. In some experiments, individual proteolytic peptide ions are isolated and subjected to fragmentation and fragment mass analysis in the mass spectrometer. The resulting fragment masses are then compared with hypothetical proteolytic peptide fragment masses of the proteins in a database. The protein is identified based on an evaluation of either or both of these comparisons.

Mass spectrometry determines a peptide mass $m_i$ to an accuracy $\pm\Delta m_i$, with $\Delta m_i/m_i$ typically >30 ppm. Within the mass range $m_i \pm \Delta m_i$ proteolytic peptide masses of several proteins in a genome database can match. Hence, an unmodified peptide will match randomly with several proteins in the database, in addition to the true match with the actual protein present in the sample, and a modified peptide will yield only random matches. Consequently, a database search using mass spectrometry information will not always identify a protein unambiguously. Therefore, in order to perform accurate and reliable molecule identification, instruments for obtaining mass data must be appropriately linked with the use of other technical resources for the comparison of mass data and mass information obtained from a database. The link can be a system that makes use of a method including means for comparison of data and database information, preferably operated via a computer.

Despite the rapidly increasing impact of mass spectrometric protein identification on proteomic research, the problem of accurately taking the phenomenon of random mass matching into account in a database search system has been overlooked. As increasingly complex processes are explored by MS-based protein identification, the use of optimized procedures will become critical. An optimized protein identification system cannot be designed without or with inappropriate account for the random mass-matching process.

## State of the Art

Identification of proteins by the above-described approach requires a scheme for determining the best match between the experimental data and a sequence in the database. Existing schemes for determining the best match include ranking by number of matches (W.J. Henzel et al., Proc. Natl. Acad. Sci. U S A 90, 5011, 1993), a scoring system based on the observed frequency of peptides from all proteins in a database in a given molecular weight range (the so-called "MOWSE score" (D.J.C. Pappin et al., Current Biology 6, 327,1993), and a scheme based on Bayesian probabilities (W. Zhang et al., Anal. Chem. 72, 2482, 2000).

None of these schemes takes the problem of random mass matching appropriately into account. The lack of an appropriate account for the random mass matching hinders optimum performance of molecule identification procedures, since the random mass matching can cause false identification results – especially when the quality of the mass spectrometry data is poor.

## Summary of the Invention

The object of the present invention is to overcome the shortcomings of the above-mentioned schemes, i.e., to provide a method that solves the problem of random mass matching.

This and other objects have been met by providing a system including methods of determining the probability for a particular score due to random mass matching of a molecule, and to utilize the computed probability to rank molecules. The method comprises: a) determining the number of matches between a database molecule and mass data; b) computing the probability that a database molecule would yield a particular number of matches by chance; c) computing a score based on one or several probabilities computed in step (b); c) comparing the scores of molecules in a molecule database; and d) identifying the molecule or molecules that yield(s) the best score(s).

The invention further provides a method of generating a frequency function of the number of matches for random (false) molecule identification for any experimental condition. The method comprises: a) defining a sub-population of the molecules contained in a database; b) computing the probability that a molecule in this sub-population would yield a particular number of matches by chance; c) computing a probability that all molecules in the sub-population would yield at most a particular number of matches by chance; d) computing the probability that at least one molecule in the sub-population would yield at least a particular number of matches by chance; and e) determining the relative frequency of each number of matches by using the probability computed in step (d) for each number of matches and generating therefrom a frequency function of the number of matches for random protein identification.

## Brief Description of the Drawings

Fig. 1 shows frequencies (i.e., number of matching proteins) of various tryptic peptide masses in a database.

Fig. 2 shows mass distribution peaks for tryptic peptides.

Fig. 3 shows the performance of an implementation of one embodiment of the invention in comparison with state of the art systems for protein identification. The graph displays results from simulations employing the invention (denoted Probity), a Bayesian method, and a method based on the number of matches.

Fig. 4 shows score frequency functions generated by the invention in comparison with score frequency functions generated by simulation.

## Detailed Description of the Invention

Many applications of molecule identification are inherently large-scale. Examples of large-scale molecule identification can be found in proteomics projects, where thousands of proteins from cells are to be identified, or cells are screened for molecular markers of states of disease. The ultimate goal of molecule identification procedures is to rely on simple, rapid and automated procedures and instrumentation. The technical solutions of the system that links and compares mass data with database information are of key importance to the design of instruments for automated molecule identification, since the system used will influence strongly the capability of obtaining a high relative frequency of true identification results, which is particularly critical when the quality of the data is poor. Furthermore, automated identification instrumentation demand that the quality of identification results is assessed automatically by the use of a significance test (J. Eriksson et al., Anal. Chem. 72, 999, 2000). However, a reliable automated protein identification system cannot be designed without or with inappropriate account for the random mass-matching process.

One object of the present invention is to provide a system that utilizes methods that allow more accurate molecule identification and more accurate and rapid significance testing of identification results. The method according to the invention appropriately takes into account the phenomenon of random matching, and is therefore well suited for implementation in an automated molecule identification system.

A particular concern regarding large-scale molecular identification is the time required to obtain the identification result together with a quality assessment of this result. A quality assessment can be accomplished by significance test, which requires knowledge of functions describing scores for false results. Such frequency functions are currently obtained by simulation of random molecular identification. However, since the time needed to derive a frequency function by simulation is about 1000 times longer than with the use of the invention, there is need to derive such a frequency function from an analytical expression. In one embodiment of the invention, such an analytical expression for the derivation of a frequency function is provided.

The methods according to the invention are well suited for, but not limited to, applications, in which the molecules are biological molecules that can exist in cells of organisms.

Biological molecules include any biological polymer that can be degraded into constituent parts. The degradation is preferably into constituent parts at predictable positions to form predictable masses. Examples of biological molecules include proteins, nucleic acid molecules, polysaccharides and carbohydrates.

An experimental biological molecule is a biological molecule that is to be identified; the experimental biological molecule can also be referred to as an unknown biological molecule. A theoretical biological molecule is a biological molecule is a known biological molecule described in a database.

Proteins are polymers of amino acids. Constituent parts of proteins comprise amino acids. A protein typically contains approximately at least ten amino acids, preferably at least 50 amino acids and more preferably at least 100 amino acids.

Nucleic acids are polymers of nucleotides. Constituent parts of nucleic acids comprise nucleotides. Typically, a nucleic acid contains at least 100 nucleotides, preferably at least 500 nucleotides.

Polysaccharides are polymers of monosaccharides. Constituent parts of polysaccharides comprise one or more monosaccharides. Typically, a polysaccharide contains at least five monosaccharides, preferably at least ten monosaccharides.

Mass data of biological molecules are quantifiable information about the masses of the constituent parts of the biological molecule. Mass data include individual mass spectra and groups of mass spectra. The mass spectra can be in the form of peptide maps, oglionucleotide maps or oligosaccharide maps.

The method of the present invention includes generating experimental mass data for the experimental molecule within a certain mass range. Mass data include the measured masses. The method also includes generating theoretical mass data in the same mass range. In one embodiment, the experimental mass data is a subset of the experimental mass data.

For example, mass data for molecules can be generated in any manner that provides mass data within certain accuracy. Examples include matrix-assisted laser desorption/ionization mass spectrometry, electrospray ionization mass spectrometry, chromatography and electrophoresis. Mass data can also be generated by a general -purpose computer configured by software or otherwise.

For the purposes of the present invention the mass data, for example a peptide mass, $m_i$, is determined to an accuracy $\pm\Delta m_i$, with $\Delta m_i/m_i$ preferably <10,000 ppm, more preferably <100 ppm, and most preferably <30 ppm.

A step in generating mass data of a molecule may include first cleaving the molecule into constituent parts. Biological molecules may be cleaved by methods known in the art. Preferably, the biological molecules are cleaved into constituent parts at predictable positions to form predictable masses. Methods of cleaving include chemical degradation of the biological molecules. Biological molecules may be degraded by contacting the biological molecule with any chemical substance.

For example, proteins may be predictably degraded into peptides by means of cyanogen bromide and enzymes, such as trypsin, endoproteinase Asp-N, V8 protease, endoproteinase Arg-C, etc. Nucleic acids may be predictably degraded into constituent parts by means of restriction endonucleases, such as Eco RI, Sma I, BamH I, Hinc II, etc. Polysaccharides may be degraded into constituent parts by means of enzymes, such as maltase, amylase, alpha-mannosidase, etc.

In the present invention a mass range ($m_{min}$, $m_{max}$) is determined for the experimental mass data. The mass range can be any mass range of the mass data. In one embodiment, the mass range is the minimum and maximum measured masses of the experimental mass data for a molecule.

A molecule database is any compilation of information about characteristics of molecules. A molecule database can be a biological molecule database. Databases are the preferred method for storing both polypeptide amino acid sequences and the nucleic acid sequences that code for these polypeptides. The databases come in a variety of different types that have advantages and disadvantages when viewed as the hypothesis for a polypeptide identification experiment.

While the "database entry" for an amino acid sequence may appear to be a simple text file for a user browsing for a particular polypeptide, many databases are organized into very flexible, complicated structures. The detailed implementation of the database on a particular system may be based on a collection of simple text files (a "flat-file" database), a collection of tables (a "relational" database), or it may be organized around concepts that stem from the idea of a protein, gene, or organism (an "object-oriented" database).

Protein mass data may be predicted from nucleic acid sequence databases. Alternatively, protein mass data may be obtained directly from protein sequence

databases that contain a collection of amino acid sequences represented by a string of single-letter or three-letter codes for the residues in a polypeptide, starting at the N-terminus of the sequence. These codes may contain nonstandard characters to indicate ambiguity at a particular site (such as "B" indicating that the residue may be "D" (aspartic acid) or "N" (asparagine)). The sequences typically have a unique number-letter combination associated with them that is used internally by the database to identify the sequence, usually referred to as the accession number for the sequence.

Databases may contain a combination of amino acid sequences, comments, literature references, and notes on known posttranslational modifications to the sequence. A database that contains these elements is referred as "annotated". Annotated databases are used if some functional or structural information is known about the mature protein, as opposed to a sequence that is known only from the translation of a stretch of nucleic acid sequence. Non-annotated databases only contain the sequence, an accession number, and a descriptive title.

The background information known about an experimental molecule by which the data base search can be constrained can include any information. Some examples of background information include information about the species of an experimental biological molecule, knowledge or an assumption about the mass of the experimental biological molecule and the isoelectric point of the experimental biological molecule.

For example, the observed molecular mass or the observed isoelectric point of a protein can be used in combination with the measured masses of peptides generated by proteolysis to constrain the search for a polypeptide. In particular, the comparison between the theoretical mass data of the database proteins and the mass data of the unknown protein may be constrained to only those proteins of the database which are within a chosen mass range. The chosen mass range is preferably within 50% of the mass of the unknown protein, more preferably within 35%, most preferably within 25%. Similarly, the comparison between the theoretical mass data of the database proteins and the mass data of the unknown protein may be constrained to only those proteins of the database which are within a chosen isoelectric point range. The isoelectric point (pI) of a protein is the pH at which its net charge is zero. The chosen isoelectric point range is preferably within 50% of the isoelectric point of the unknown protein, more

preferably within 35%, most preferably within 25%.

Optionally, further information of the experimental biological molecule, such as a protein's sequence, is obtained by generating fragment mass data of the experimental and theoretical biological molecules. Fragment mass data for a peptide can be generated in any manner which provides fragment mass data within a certain accuracy. Experimental conditions include the type of energy used to generate the fragment mass data. Vibrational excitation energy can be used. The vibrational excitation may be generated by collisions of the peptide with electrons, photons, gas molecules or a surface. Electronic excitation can be used. The electronic excitation may be generated by collisions of the peptide with electrons, photons, gas molecules (e.g. argon) or a surface.

In another example, the experimental fragment mass spectrum of a peptide from an enzymatically digested unknown protein is compared with the theoretical masses calculated by applying the rules for the specificity of the enzyme, and the rules for the fragmentation as known to those of ordinary skill in the art, to the amino acid sequence of a database protein.

Fragment mass data for the purposes of this invention can be generated by using multidimensional mass spectrometry (MS/MS), also known as tandem mass spectrometry. A number of types of mass spectrometers can be used including a triple-quadruple mass spectrometer, a Fourier-transform cyclotron resonance mass spectrometer, a tandem time-of-flight mass spectrometer, and a quadruple ion trap mass spectrometer. A single peptide from a protein digest is subjected to MS/MS measurement and the observed pattern of fragment ions is compared to the patterns of fragment ions predicted from database sequences.

In one embodiment, the invention provides a method to determine the probabilities for the scores that a particular molecule in a database can yield by chance when compared with mass data. The method can operate under a variety of experimental and database search constraints. The score can be the number of matches between masses derived from known or hypothetical molecules or molecular constituents in a database and masses in mass data from one or several known or unknown molecules, or molecular constituents. The score can also result from a computation that utilizes the number of matches.

In one embodiment, the invention provides a method to extract information about the molecules in a database. Examples of information that can be extracted from a database are total molecular mass, charge, isoelectric point, hydrophobicity and known or hypothetical chemical modification, and mass,

charge, isoelectric point, hydrophobicity and known or hypothetical chemical modification of molecular constituents.

In one embodiment, the invention provides a method to perform actions on molecules in the database that are supposed to mimic actions occurring in an experiment. Examples of actions are degradation of molecules into molecular constituents by hydrolysis, where hydrolysis can result from the activity of chemicals or enzymes. The method can also perform actions that mimic experimental actions on molecular constituents. For example, the fragmentation of an excited molecular constituent into smaller pieces.

In one embodiment, the invention provides a method to derive a number of molecular pieces, $k_u$, resulting from an action assumed to mimic an experimental situation. The pieces can be molecular constituents, such as proteolytic peptides resulting from enzymatic digestion of a protein, where different assumptions can be made concerning the degree of completeness of the enzymatic digestion. The pieces can be molecular constituents in the form of fragments of molecular constituents, e.g. fragments of proteolytic peptides.

In one embodiment, the invention provides a method to organize the masses of molecules or molecular constituents or fragments thereof. Examples of such organization are given in Fig. 1 and 2, where Fig. 1 displays the number of proteins in a database that match a given proteolytic peptide mass and Fig 2 displays the clustered distribution of proteolytic peptide masses. Masses clustering in this or similar fashions will be referred to as a mass distribution peak. Mass distribution peaks can be found for all molecules that contain a limited number of different atoms (e.g. C, H, N, O, S).

In one embodiment, the invention provides a method for defining mass regions wherein the frequency of various masses can be determined. The method defines $f_i$ as the fraction of masses of molecular constituents or fragments that falls into a mass region $i$.

In one embodiment, the invention provides a method that determines a probability $p_i$ that a particular molecule in a database will be found in a randomly chosen mass distribution peak in the mass region $i$:

$$p_i = F(k_u, m_i, c),$$

where $F$ is a function, $m_i$ is a mass region, and $c$ denotes experimental and database search constraints.

In one embodiment $p_i$ is given by:

$$p_i = f_i \cdot \frac{k_u}{m_{i+1} - m_i},$$

which describes the probability that a molecular constituent from a particular molecule characterized by $k_u$ will be found in a single randomly chosen mass distribution peak. The denominator of the expression above describing $p_i$ represents the number of mass distributions peaks within the mass region $i$.

In one embodiment the invention provides a method of determining the probability, $p_i'$, of finding a molecular constituent originating from a particular molecule characterized by $k_u$ within a region $\pm\Delta m$ around a randomly chosen molecular constituent mass $m$:

$$p_i' = p_i \cdot \delta(m_i, \Delta m),$$

where $\delta(m_i, \Delta m)$ denotes a function that depends on the shape of the mass distribution peak and $m_i$ refers to a mass region. $\delta(m_i, \Delta m)$ can be interpreted as a statistical measure of the number of molecular constituent masses that can be found within $\pm\Delta m$ from a randomly chosen molecular constituent mass. The mass accuracy $\Delta m$ can be different for different mass regions, i.e., in that case denoted by $\Delta m_i$.

In one embodiment, the invention provides a method to determine $\delta(m_i, \Delta m)$ by simulation of the relative frequency of masses around a randomly chosen mass in a mass distribution. In one embodiment, $\delta(m_i, \Delta m)$ is determined by integration of a function describing molecular constituent mass distributions and normalization to the total number of molecular constituent masses in a mass distribution peak. In one embodiment, $\delta(m_i, \Delta m)$ is determined by direct counting followed by normalization.

In one embodiment of the invention, a finite number of mass regions between $m_{min}$ and $m_{max}$ is employed, each having an individually defined $p_i'$.

In one embodiment the probabilities $p_i'$ are employed to compute a total probability, $p(k)$, for an individual molecule in the database to match randomly $k$ out of $n$ masses, where the $n$ masses refers to the number of masses in the mass data.

$$p(k) = G(p_i', k, n, c'),$$

where $G$ is a function and $c'$ denotes experimental and database search constraints.

In one embodiment of the invention $p(k)$ is given by:

$$p(k)=\sum_{k_i,\sum k_i=k}\{\frac{n_1!}{k_1!(n_1-k_1)!}\cdot p_1'^{k_1}\cdot\left(1-p_1'\right)^{n_1-k_1}\cdot\frac{n_2!}{k_2!(n_2-k_2)!}\cdot p_2'^{k_2}\cdot\left(1-p_2'\right)^{n_2-k_2}\cdot\ldots$$
$$\cdot\frac{n_q!}{k_q!(n_q-k_q)!}\cdot p_q'^{k_q}\cdot\left(1-p_q'\right)^{n_q-k_q}\}$$

where $q$ denotes the number of mass regions, $n_1$ denotes the number of masses in the mass data that are in mass region 1, $n_2$ denotes the number of masses in the mass data that are in mass region 2 etc., and $k_i$, where $i=1,2,\ldots,q$, denotes the number of matches in mass region $i$. The values of $k_i$ are all combinations of values that apply to the constraint $\sum k_l = k$.

In one embodiment of the invention, a score related to random matching is employed in the process of ranking molecules in a database.

In one embodiment of the invention, the probability $p(k)$ is employed in the process of ranking molecules in a database. A whole database or a fraction of a database is processed and organized to allow the computation of $p(k)$ for molecules in the database. $k$ denotes the number of matches between the masses of molecular constituents of each database molecule investigated and masses in the mass data. The molecules in the database can be known or hypothetical. The molecule or molecules producing the mass data can be known or unknown.

In one embodiment of the invention, the ranking of the molecules in a database is based on the score $S(p(k))$, where $S$ is a function.

In one embodiment of the invention

$$S(p(k))=c\cdot(1-\sum_{k'<k}p(k'))=c\sum_{k}^{n}p(k'),$$

where $c$ is a constant or a mathematical function. When $c$=1, S(p(k)) can be interpreted as the probability that a molecule in the database would yield at least $k$ random matches with the mass data.

In one embodiment of the invention, the molecule in the database that yields the lowest $S(p(k))$ for $k$ matches with the mass data is given the highest rank. The molecule in the database yielding the second lowest $S(p(k))$ for $k$ matches is given the second highest rank and so on. The identification of a molecule or molecules is among the molecules having the highest ranks. The highest ranks can be the top ranked molecule only, but it can also be more molecules than the top ranked, e.g. the top two, top three, top four, top five, top ten, or top 100. The number of ranked molecules that are considered as identification results can also be determined by the use of a significance test.

In one embodiment, the invention provides a method of generating a frequency distribution of scores for a particular experimental condition, wherein the scores relate to random identifications of proteins.

A frequency distribution is any compilation of the observed values of the variable being studied and how many times each value is observed. Frequency distributions can be in the form of a table of listings, a bar graph, a histogram, a frequency polygon, or a continuous curve. Functions derived from frequency distributions can be continuous (probability density function) or discrete (probability mass functions). Cumulative distribution functions of each type of function can also be derived.

In one embodiment, the frequency function is generated for a sub-population with $H$ members from a database.

In one embodiment, the sub-population is selected based upon values of $k_u$.

In one embodiment, the frequency function is generated for molecules ranked upon their number of matches.

In one embodiment, the frequency function is $f(S)$, where $S$ is a score. In one embodiment, $S$ is the number of random matches.

In one embodiment $S=k'$ and

$$f(S)=\left\{\sum_{k=0}^{k'} p(k)\right\}^{H}-\left\{\sum_{k=0}^{k'-1} p(k)\right\}^{H},$$

where $p(k)$ has the meaning stated above.

Those of ordinary skill in the art will recognize that the present invention has wide applicability for identification of molecules. Although illustrative embodiments of the present invention have been described herein with reference to the accompanying drawings, it is to be understood that the invention is not limited to those precise embodiments, and that various other changes and modifications may be effected therein by one skilled in the art without departing from the scope or spirit of the present invention.

## Claims

1. A method of assigning an identity to one or several different molecules in a sample by comparison of characteristics obtained under certain conditions for said sample with stored characteristics of individual ("stored") molecules, which method is **characterized by** the steps of:

a) determining the number, $k$, of matches between stored characteristics of said individual molecules and characteristics observed from the sample;

b) computing the probability, $p(k)$, that a particular one of said stored individual molecules has characteristics that match randomly the characteristics of the sample;

c) assigning an individual score, $S(p(k))$, for a number of said stored molecules based on number of matches determined in step (a) and the probability computed in step (b);

d) ranking each of the individual stored molecules that in step (c) have been assigned an individual score according to this score; and

e) assigning an identity to one or several molecules, the characteristics of which were obtained under certain conditions based on the ranking in step (d).

2. The method according to claim 1, further **characterized in** that the determination of the number of matches in the step (a) of determining the number of matches in claim 1 is between characteristics of stored molecules assuming that these molecules have been subjected to the same conditions as the molecules in the sample.

3. The method according to claim 1 or 2, further **characterized in** that said characteristics are masses of the constituents of the stored molecules, which masses cluster in mass distribution peaks, and that the step (b) of computing a probability in claim 1 comprises the steps of:

a) determining the masses and the number, $k_u$, of masses that can be generated for the particular condition for each individual molecule of the stored molecules;

b) defining a total number, $q$, regions, $i$, of the masses that have been computed in step (a);

c) determining a fraction, $f_i$, of all the masses computed in step (a) that are within a region $i$ as defined by step (b);

d) calculating a probability, $p_i' = f_i \cdot \frac{k_u}{m_{i+1} - m_i} \cdot \delta(m_i, \Delta m)$, where the denominator is the number of mass distribution peaks in the mass region $i$ defined in step (b) above, and $\delta(m_i,\Delta m)$ is a statistical measure of the number of constituent masses that can be found within $\pm\Delta m$ from a randomly chosen molecular constituent mass, which means that $p'i$ is the probability of finding a molecular constituent originating from a particular stored molecule within a region $\pm\Delta m$ around a randomly chosen constituent mass;

e) determining the probabilities as described in step (d) for all regions defined in step (b);

f) determining the number, $n_i$, of masses in the mass data that fall into each of the $q$ the mass regions $i$ defined by step (b); and

g) determining the probability

$$p(k) = \sum_{k_1, \sum k_i = k} \{ \frac{n_1!}{k_1!(n_1 - k_1)!} \cdot p_1'^{k_1} \cdot \left(1 - p_1'\right)^{n_1 - k_1} \cdot \frac{n_2!}{k_2!(n_2 - k_2)!} \cdot p_2'^{k_2} \cdot \left(1 - p_2'\right)^{n_2 - k_2} \cdot \ldots$$
$$\cdot \frac{n_q!}{k_q!(n_q - k_q)!} \cdot p_q'^{k_q} \cdot \left(1 - p_q'\right)^{n_q - k_q} \}$$

for a particular individual stored molecule to match randomly $k$ out of $n$ masses, where the $n$ masses refers to the number of masses in the mass data.

4. The method according to anyone of claims 1 to 3, **characterized in** that said characteristics are masses of the constituents of the stored molecules, which masses cluster in mass distribution peaks, and that the step (c) of assigning an individual score in claim 1 comprises the step of calculating the score according to $S(p(k)) = c \cdot (1 - \sum_{k'<k} p(k'))$, where c is a constant or a function or operator.

5. The method according to anyone of claims 1 to 4, **characterized in** that said molecules are biological molecules.

6. The method according to anyone of claims 3 to 5, **characterized in** that said masses are obtained with mass spectrometry.

7. A method to determine a frequency function, *f(S)*, of random molecule identification based on the method of computing the probability *p(k)* according to claim 1, which method is **characterized by** the steps of:

a) defining a sub-population, with $H$ members, of the stored molecules; and

b) calculating the frequency function according to

$$f(S)=\left\{\sum_{k=0}^{k'} p(k)\right\}^{H}-\left\{\sum_{k=0}^{k'-1} p(k)\right\}^{H}, \text{ where } S=k'.$$

8. A method to determine a frequency function, *f(S)*, of random molecule identification based on the method of computing the probability *p(k)* according to claim 3, which method is **characterized by** the steps of:

c) defining a sub-population, with $H$ members, of the stored molecules where the members of the sub-population are selected based on their values of $k_u$; and

d) calculating the frequency function according to

$$f(S)=\left\{\sum_{k=0}^{k'} p(k)\right\}^{H}-\left\{\sum_{k=0}^{k'-1} p(k)\right\}^{H}, \text{ where } S=k'.$$

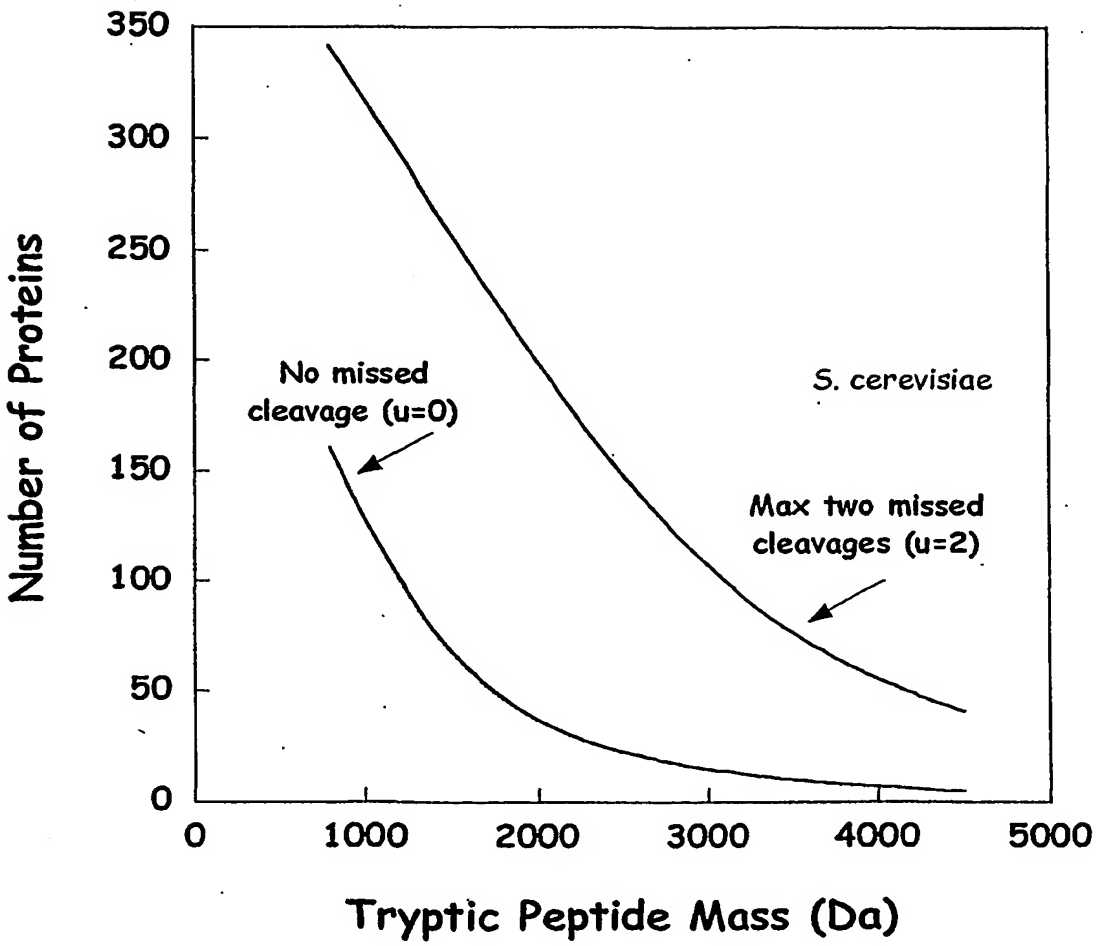

Number of Proteins
350
300
250
200
150
100
50
0
0
1000
2000
3000
4000
5000
Tryptic Peptide Mass (Da)
No missed cleavage (u=0)
Max two missed cleavages (u=2)
S. cerevisiae

Fig. 1

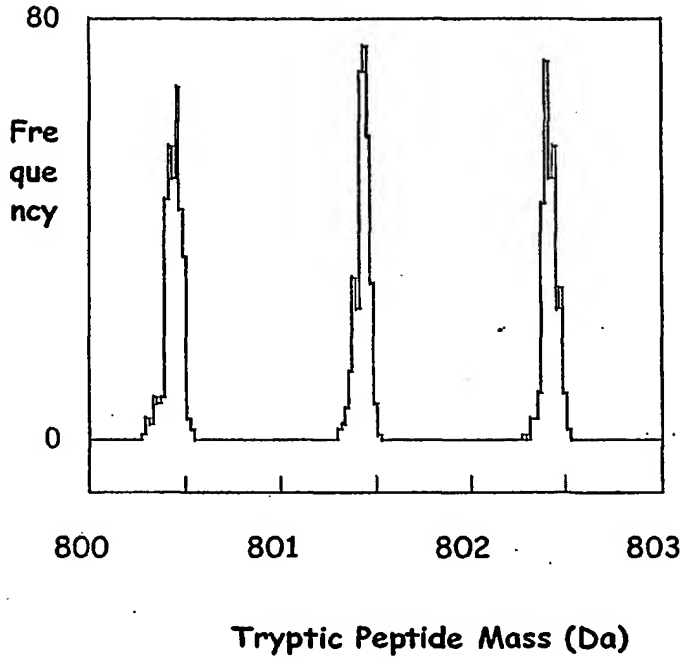



**Fig. 2.**

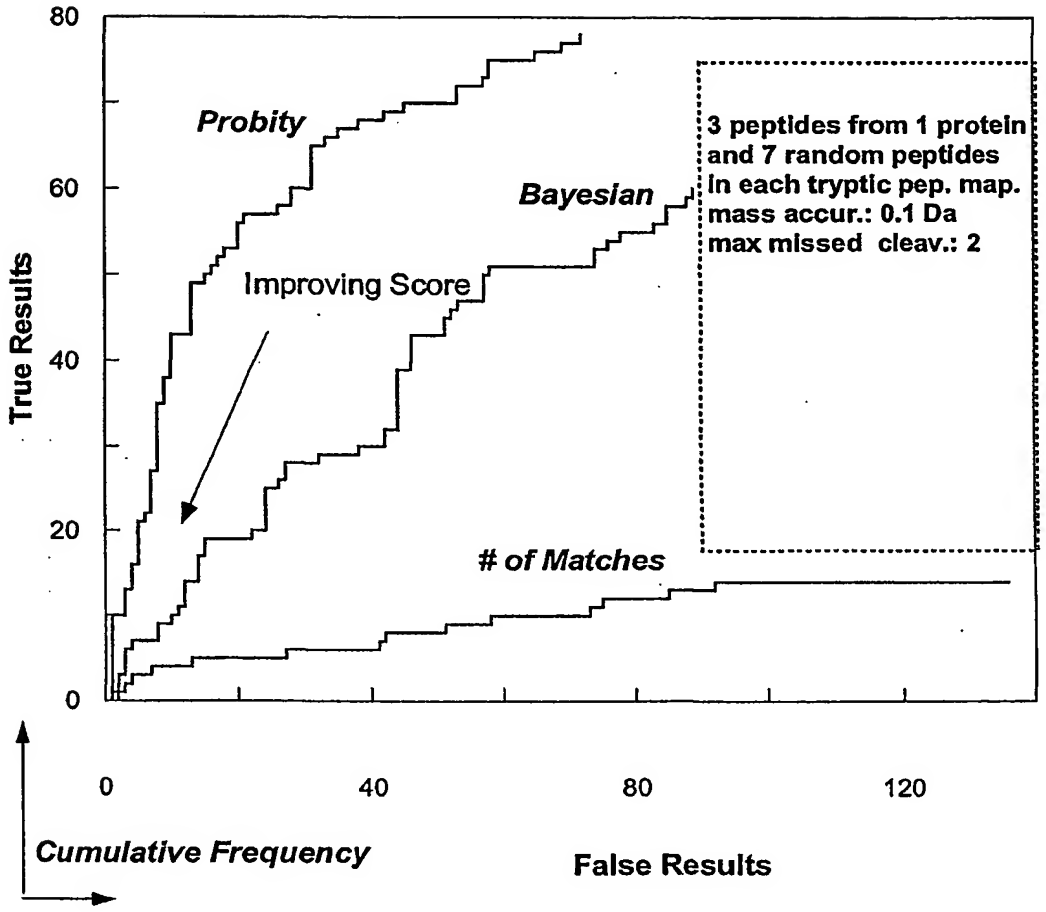

80
60
40
20
0
True Results
Probity
Bayesian
Improving Score
# of Matches
3 peptides from 1 protein
and 7 random peptides
In each tryptic pep. map.
mass accur.: 0.1 Da
max missed cleav.: 2
0
40
80
120
Cumulative Frequency
False Results

Fig. 3

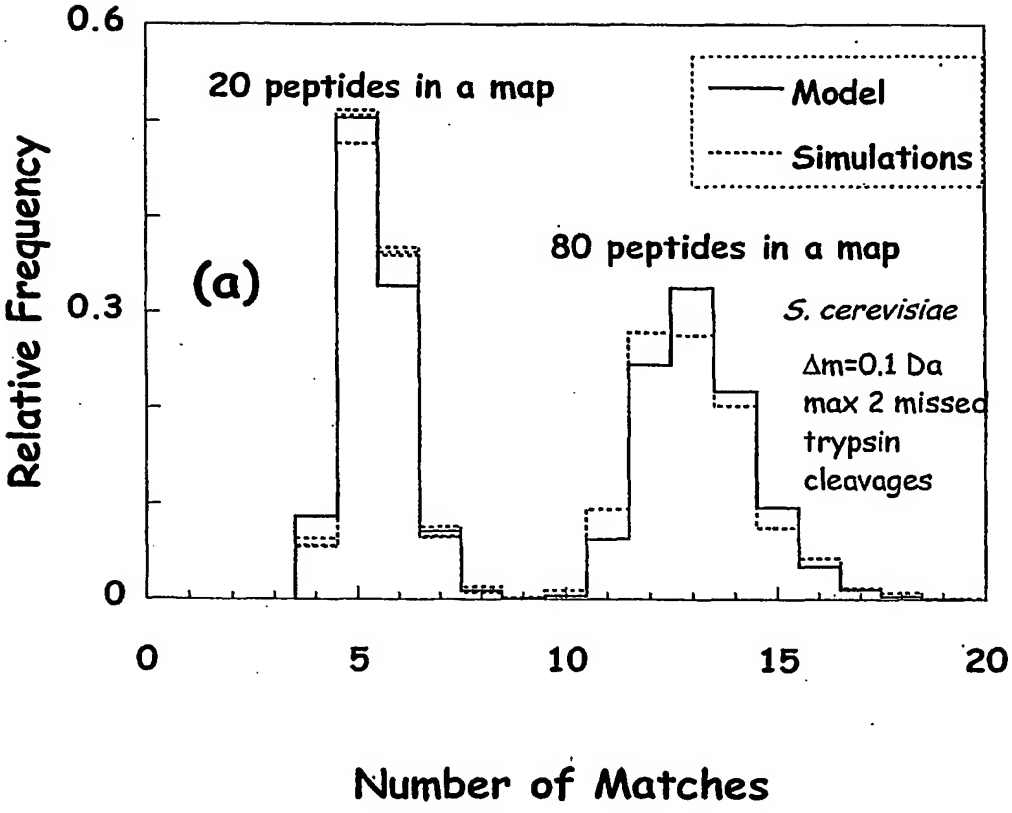

0.6
0.3
0
Relative Frequency
20 peptides in a map
Model
Simulations
(a)
80 peptides in a map
S. cerevisiae
Δm=0.1 Da
max 2 missed
trypsin
cleavages
0
5
10
15
20
Number of Matches

Fig. 4

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/SE 01/01322

A. CLASSIFICATION OF SUBJECT MATTER

IPC7: G01N 33/00, H01J 49/26, H01J 49/40
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC7: G01N, H01J

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

SE,DK,FI,NO classes as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| P,A | WO 0073787 A1 (ROCKEFELLER UNIVERSITY ET AL), 7 December 2000 (07.12.00), page 2, line 10 - line 22; page 8, line 20 - page 9, line 21; page 10, line 26 - page 12, line 25, figures 1,3, claims 1-43, abstract<br>-- | 1-8 |
| A | Patent Abstracts of Japan, abstract of JP -48765 A (JEOL LTD), 18 February 2000 (18.02.00), figure 1, abstract<br>-- | 1-8 |
| P,A | EP 1047107 A2 (MICROMASS LIMITED), 25 October 2000 (25.10.00), page 3, line 30 - page 4, line 42, claims 1-3, abstract<br>-- | 1-8 |

[X] Further documents are listed in the continuation of Box C. [X] See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance: the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| 25 October 2001 | 07-11-2001 |
| Name and mailing address of the ISA/<br>Swedish Patent Office<br>Box 5055, S-102 42 STOCKHOLM<br>Facsimile No. +46 8 666 02 86 | Authorized officer<br>Klas Arvidsson / MRo<br>Telephone No. +46 8 782 25 00 |

Form PCT/ISA/210 (second sheet) (July 1998)

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/SE 01/01322

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| A | US 5538897 A (J.R. YATES, III ET AL), 23 July 1996 (23.07.96), column 1, line 65 - column 2, line 10; column 2, line 45 - column 5, line 57, figure 3, claims 1-19, abstract<br><br>--<br>-------- | 1-8 |

# INTERNATIONAL SEARCH REPORT

Information on patent family members

01/10/01

International application No.
PCT/SE 01/01322

| Patent document cited in search report | Publication date | Patent family member(s) | Publication date |
|---|---|---|---|
| WO 0073787 A1 | 07/12/00 | NONE | |
| EP 1047107 A2 | 25/10/00 | EP 1047108 A | 25/10/00 |
| | | GB 9907810 D | 00/00/00 |
| | | GB 9908684 D | 00/00/00 |
| US 5538897 A | 23/07/96 | CA 2185574 A | 21/09/95 |
| | | EP 0750747 A | 02/01/97 |
| | | JP 9510780 T | 28/10/97 |
| | | US 6017693 A | 25/01/00 |
| | | WO 9525281 A | 21/09/95 |